In Large Language Models (*LLMs*), changing sampling *temperature* *does not* affect performance on *problem-solving* tasks*.

* For temperatures from 0.0 to 1.0 on multiple-choice question-answer (MCQA) problems.

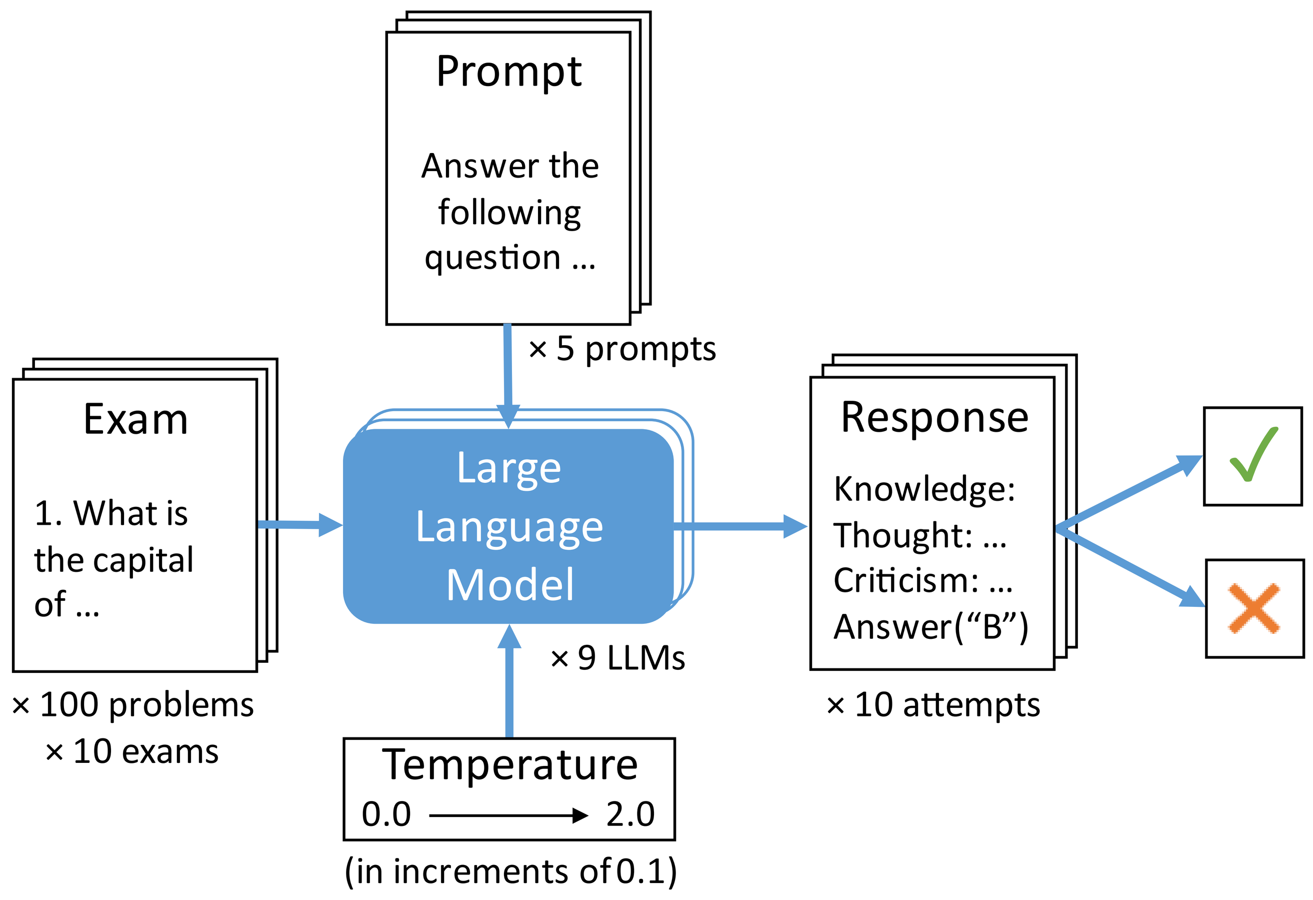
Learn more

Question

What is the optimal sampling temperature for LLM performance on problem-solving tasks?

Experiment

To measure the effect of sampling temperature on problem solving, we used 9 LLMs, with 5 prompts, to solve 1,000 problems, from 10 exams, 10 times each, across temperatures of 0.0 to 2.0 in increments of 0.1.



Models

Name	Vendor
Claude 3 Opus	Anthropic
Command R+	Cohere
Gemini 1.0 Pro	Google
Gemini 1.5 Pro	Google
GPT-3.5 Turbo	OpenAI
GPT-4	OpenAI
Llama 2 7B	Meta
Llama 2 70B	Meta
Mistral Large	Mistral AI

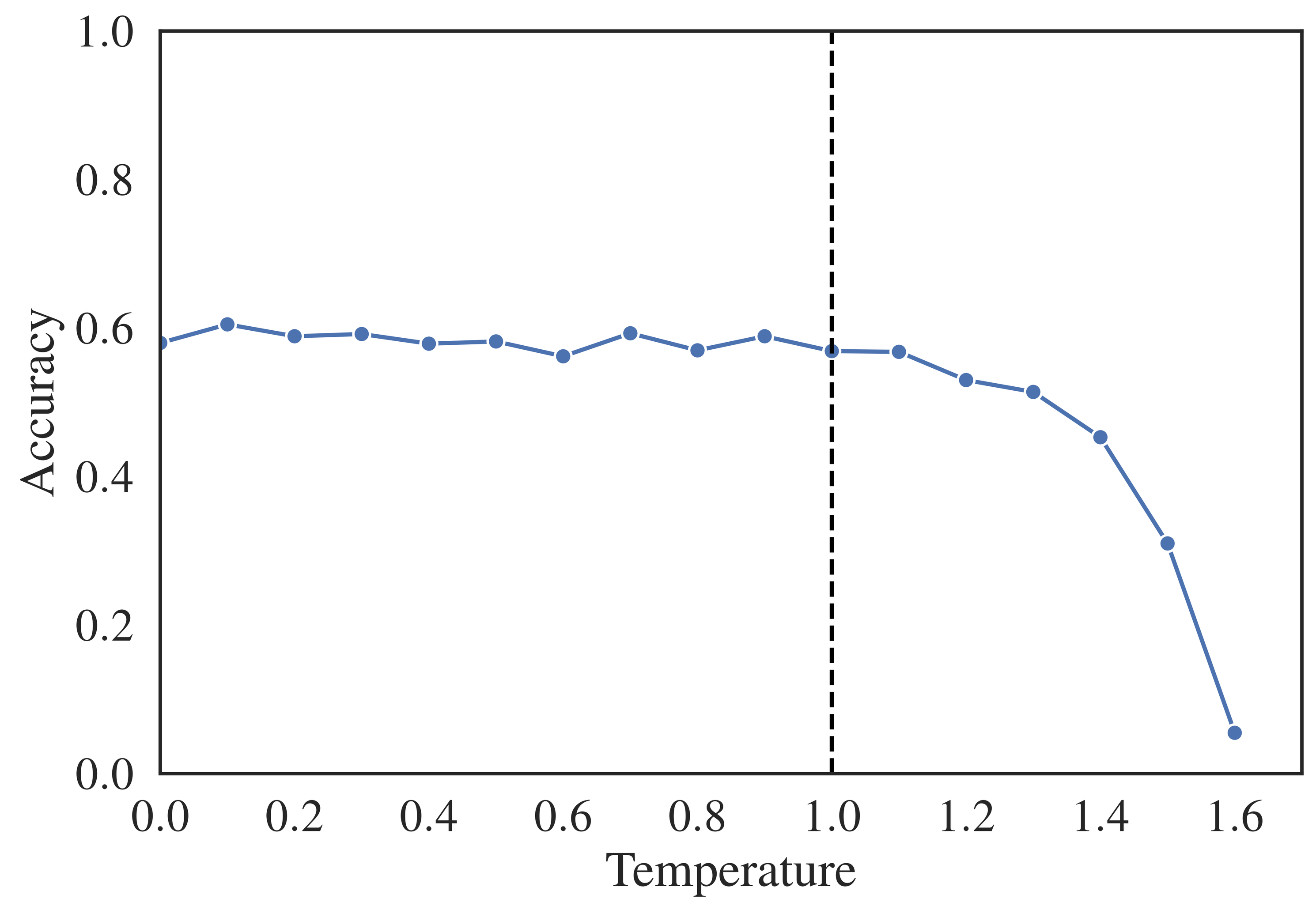
Problem Sets

Name	Benchmark
Arc Challenge	ARC
AQUA-RAT	AGI Eval
Hellaswag	HellaSwag
LogiQA	AGI Eval
LSAT-AR	AGI Eval
LSAT-LR	AGI Eval
LSAT-RC	AGI Eval
MedMCQA	MedMCQA
SAT-English	AGI Eval
SAT-Math	AGI Eval

Prompts

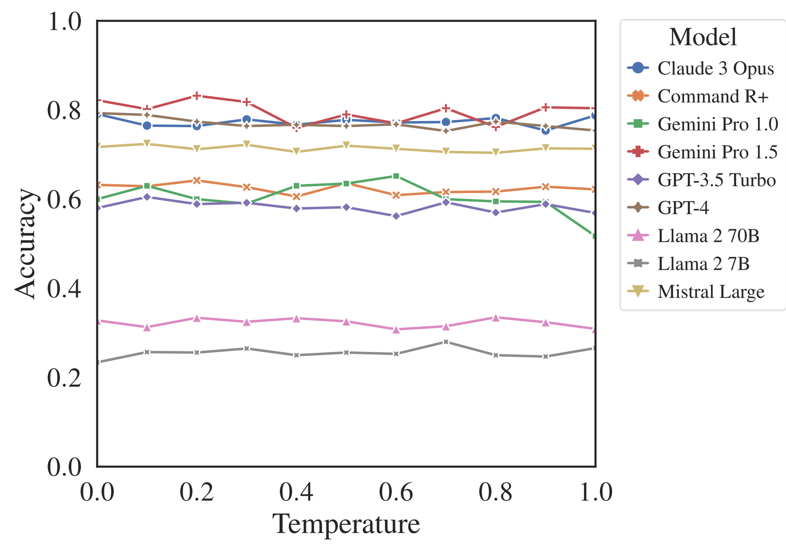
Baseline – no prompt engineering (used as a baseline)
Domain Expertise – specifies the LLM is an expert in the problem domain
Self-recitation – instructs the LLM to recite its own internal knowledge first
Chain of Thought – instructs the LMM to "think step-by-step"
Composite – combines all three prompts and adds self-criticism

Results

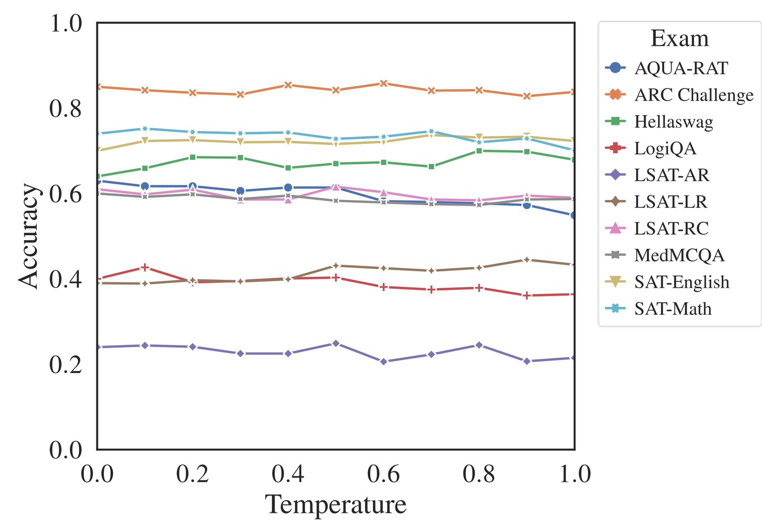


Accuracy by temperature from 0.0 to 1.6 for GPT-3.5 using CoT prompt on the 100-question exam.

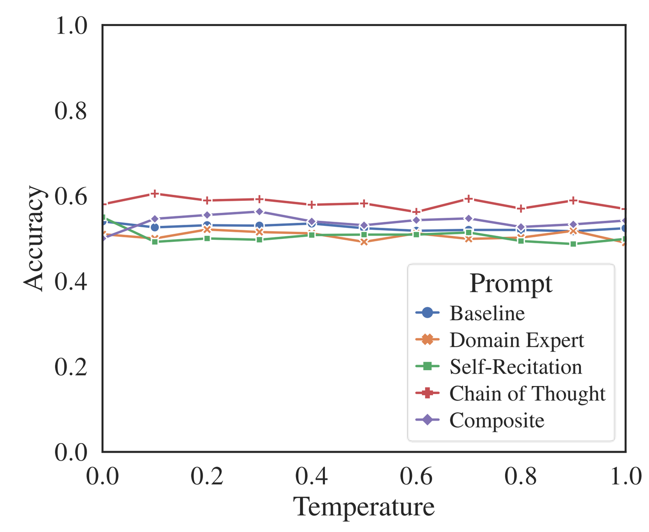
More Results



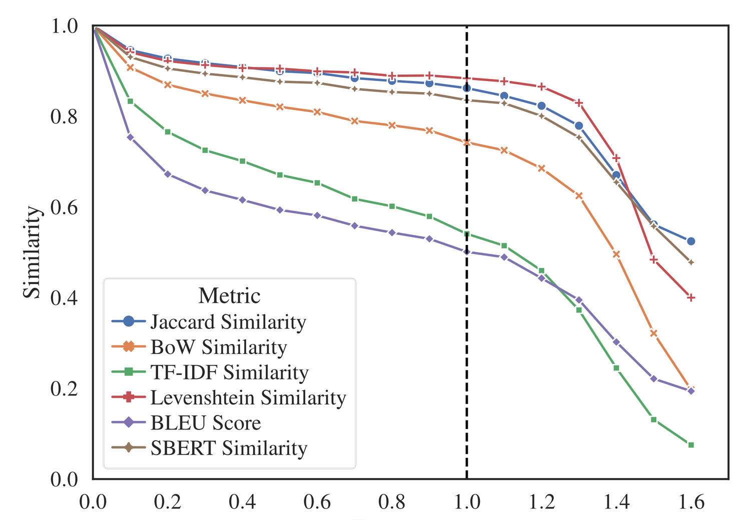
Accuracy by temperature and model using the CoT prompt on the 100-question exam.



Accuracy by temperature and exam for GPT-3.5 using the CoT prompt.



Accuracy by temperature and prompt for GPT-3.5 using the CoT prompt on the 100-question exam.



Text similarity by temperature for GPT-3.5 using the CoT prompt on the 100-question exam.

Conclusion

Changes to sampling temperature from 0.0 to 1.0 have no statistically significant effects on problem-solving for multiple-choice question-answer (MCQA) problems.

The Effect of Sampling Temperature on Problem Solving in Large Language Models

Matthew Renze and Erhan Guven
Johns Hopkins University